chinaXiv:202304.00947v1

基于知识元引用网络的细分领域演化特征研究*

毛进 [1,2]，侯博文 [1,2]，王依蒙 [2]
1 武汉大学信息资源研究中心　武汉　430072
2 武汉大学信息管理学院　武汉　430072

摘要：[目的/意义]理解科学知识发展演化过程有助于促进科学研究，从微观视角追踪知识在细分领域中的结构与演化特征对知识评估和知识服务具有重要意义。[方法/过程]以医学信息学中知识元为例，利用语义类型界定每种疾病的治疗相关细分领域，构建 125 种疾病不同时刻的知识元引用网络，采用 Leiden 算法识别知识群落，从群落知识演化、群落知识竞争状态等维度揭示疾病个体的演化特征；提出丰富度、均衡性与差异度三种测度指标，揭示疾病个体与整体的多样性特征。[结果/结论]研究表明，知识元引用网络中的群落能够反映出疾病知识结构与演进状态。整体层次的疾病知识多样性特征包括：疾病知识群落数量不断增加，群落间规模与组成差异不断扩大；不同疾病表现出常规型、早期争议型、泛指型演化模式；研究较早的疾病普遍均衡性较低、差异度较高。

关键词：SPO 三元组　知识元　知识群落　演化特征　知识多样性
分类号：G203

引言

随着全球科学出版物的急速增长，挖掘和分析科学知识的演进特征与规律具有重要意义，揭示科学知识结构及其演化规律日益成为图书情报学科重要的研究问题。领域与细分领域是科学知识在不同层次的聚合，领域中往往存在多种不同的主题，而细分领域则表示领域中某一主题下的同类型知识内涵。例如，若将生物医学学科中与治疗相关的所有科学知识视为一个研究领域，那么其中糖尿病和哮喘有关的疾病治疗知识则可以分别视为一个细分领域。当前科学知识相关研究较多关注于学科或者研究领域层次，而分析细分领域知识结构并揭示其特征将有助于更加深入地理解科学知识发展过程。

从分析对象来看，相关研究多采用关键词或主题词来进行领域主题的挖掘与分析，然而关键词或主题词仅仅是文献表层知识特征的反映[1]。知识元是一种从科学文献内容中解析出的知识单元，能更加细粒度地反映科学知识的内部结构。SPO 三元组是由主语（Subject）、谓语（Predication）和宾语（Object）组成的表征一定语义内容和关系的知识元三元组。相关研究较多关注基于三元组的药物治疗[2]、基因诊断[3]以及其他知识图谱[4]相关研究，忽视了三元组在知识演化分析中的重要作用。实际上作为一种可操作的表示知识元的数据模型，三元组能够在

*本文系国家自然科学基金项目“基于‘问题-方法’关联识别的科学知识创新探测与协同演化分析”（项目编号：72174154）和国家自然科学基金创新研究群体项目“信息资源管理”（项目编号：71921002）的研究成果之一。

作者简介：毛进(ORCID：0000-0001-9572-6709)，副教授，博士，博士生导师，E-mail: danveno@163.com；侯博文(ORCID：0000-0001-6541-8430)，硕士研究生；王依蒙(ORCID：0000-0002-1645-7502)，硕士研究生。

文献引用过程中发生承接关系，是知识计量的一种可行途径[5]。不同于关键词或主题词的过度抽象凝练，三元组丰富的语义类型能够更为精准地界定知识所属的细分领域。另一方面，引文网络中三元组所组成的社群也可以更为全面地反映细分领域知识组成。

由此，本文提出了一种针对 SPO 三元组的细分领域演化分析方法，基于知识元引用网络识别细分领域的知识群落，进而分析细分领域的多样性与演化特征，最后挖掘整体领域的多样性演化规律与模式。本文首次提出利用知识元三元组构建引用网络，通过三元组的语义类型与知识群落兼顾了科学知识的完整性和结构化，为分析知识演化的内容与路径等特征提供了新的视角。同时，本文利用知识群落测度细分领域的多样性，归纳出潜在的科学知识演化模式，能够为领域知识多样性评价提供参考借鉴，丰富知识元相关理论研究。

1 相关研究

1.1 细粒度科学知识表示

科学知识的细粒度表示是发现其结构与深层特征的前提，知识元是当前细粒度科学知识表示的重要形式之一，其与三元组和知识模因有着密不可分的联系。最早有学者认为知识元是一种由向导信息和知识组成的知识结构[6]，后续知识元内涵被重新界定为“N 个语义三元组的逻辑组合”，并据此形成知识元的 SPO 三元组描述模型[7]。当前有研究通过从海量 SPO 三元组中挖掘矛盾性知识[8]或是针对某一类疾病发现规律性语义模式[9]进行一定的知识发现。知识模因则是引文关系或共现关系中伴随主题演化的稳定性术语[10]，其往往是在引文关系中获得继承和复制的简短的文本单元[11]，有研究将其用于分析跨学科领域的学科结构[12]。因而三元组中的主语、谓语和宾语作为具有实际语义的知识模因能够界定领域与细分领域，成为知识元引用网络演化分析的重要切入点。

知识演化与知识群落的相关研究源于对类似生物问题的借鉴。波普尔首次从生物进化的视角看待知识的生长发展问题[13]，后续学者陆续丰富知识进化理论，认为知识进化的演变发展遵循着知识的遗传和变异机制等规律[14]。知识的遗传机制是指知识通过继承与传递实现了知识的延续性；知识的变异机制是指不同的知识片段重新组合成新的知识（基因）的过程。类比于生物群落的概念模型，知识群落被界定为基于知识间潜在的内在联系或特定目标而形成的，具备一定生物属性的知识群集[15]，并且知识间的引用关系能够塑造领域知识的群落结构[16]。在图书情报学领域，针对知识群落的分析与复杂网络理论中的派系分析方法较为类似，并逐步应用于探索知识发展脉络及规律[17]。受此启发，本文对知识元引用网络开展知识群落层面的演化研究与多样性分析，以期弥补以往相关研究的缺失。

1.2 领域知识演化

主题演化揭示的结果往往比知识演化过程更为宏观[18]，前者更适用于不同主题的演变，在知识演化相关问题上的可移植性较差。部分研究针对具体研究领域识别演进主题，并构建演进路径进行分析。例如对医学专利数据的 SAO 结果进行词频统计并构建语义网络识别核心技术主题与发展阶段[19]、对美国国家科学基金会（NSF）数据中 AI 领域标题与摘要提取关键词并进行主题挖掘[20]，或是直接将生物医学领域的 SPO 三元组作为文献的替代，以构建基于谓词的语义网络并识别新兴研究主题[21]。更多研究关注构建关键词的语义网络[22]、共现网络[23]或是主题词的关联网络[24]，并结合 LDA 模型[25]挖掘主题。演进路径的构建则基

chinaXiv:202304.00947v1

于不同时间窗口主题之间的相似性，例如有学者从共词网络出发，利用关键词的共现赋予权重并构建距离矩阵确定演进路径[26]。

不同于主题演化，知识往往通过引用关系在某一主题内演变，众多学者围绕知识模因引文网络分析知识演化。有研究从知识趋同、知识聚合与发散、主题流动（Topic dynamics）三个角度挖掘知识在引文间的联系情况[27]，或是利用知识生命周期理论总结知识演进过程，通过关键词对的直接与间接两类引用方式构建知识演化路径[28]。除此之外，也有研究利用引用关系构建知识模因的扩散级联网络并发现医学信息学的四种扩散模式[29]，或是基于知识基因流动和扩散两种机制分析引文间知识的遗传和变异情况[30]。部分学者开始通过知识元开展知识演化研究，例如对不同时期的 ESI 前沿领域知识元集合测度其迁移与重组情况以展现微观知识演进过程与规律[31]，但整体疏于描述知识结构，而细粒度的三元组模型则为开展细分领域知识演化提供了一条可行的路径。

2 基于知识元引用网络的知识群落发现

2.1 细分领域的知识元引用网络构建

知识元引用网络是指将处于相同细分领域的知识元作为节点，以知识元所属文献间引用关系为边，构成的知识网络。整体领域由谓语所表述的语义限定，细分领域则由知识元的主语以及宾语的语义类型限定（如图 1）。以三元组“卡巴拉汀-治疗-阿尔兹海默症”为例，首先通过谓词“治疗”确定该三元组所属的治疗领域，进而由药物“卡巴拉汀”与疾病“阿尔兹海默症”将该知识元节点归属于治疗领域内以阿尔兹海默症为细分领域的知识元网络。截至时间T_i组成的网络形式化表示为$G_{Ti}=\{N_{Ti},E_{Ti}\}$，其中N为知识元节点集合，E为有向边集合。

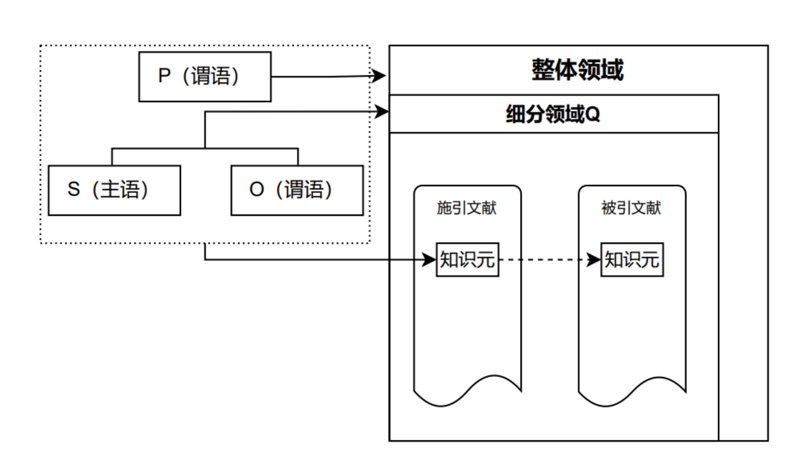


图 1 知识元引用网络构建示例

按照知识元引用网络定义，描述具体构建过程：

Step1. 初始化知识元引用网络G={N,E}，选定文献集 P 与细分领域 Q，领域由文献集 P 中所抽取的 SPO 三元组的谓语决定，细分领域由 SPO 三元组中的主语以及宾语的语义类型决定。

Step2. 选择细分领域 Q 内所有 SPO 三元组作为节点，得到节点集合$N=\{n_1,n_2,\dots,n_i\}$

Step3. 以节点集合所属文献间的引用关系作为边得到边集合$E=\{e_1,e_2,\dots,e_i\}$

Step4. 按照文献集 P 出版时间划分不同时间 Ti 的网络集合$G_{Ti}=\{N_{Ti},E_{Ti}\}$

chinaXiv:202304.00947v1

Step5. 输出细分领域动态知识元引用网络集$W=\{G_{T1},G_{T2},\ldots,G_{Ti}\}$

2.2 基于 Leiden 算法的群落识别

Leiden 算法是一种分层聚类算法，其修改了 Louvain 算法中部分社区连接不紧密的问题，通过节点的局部移动、分区细化与网络聚合实现了良好的社区识别效果[32]。在知识元引用网络中，知识群落是指聚集相似知识并通过引用关系连接在一起的知识元语义集合。划分不同时刻的网络社区，能够识别细分领域内的知识群落并呈现知识的扩散与转移情况。本文运用 Leiden 算法从知识元引用网络中识别社区，由于结构过小的社区可能不足以形成成熟的知识群落，因此本文将知识群落的节点阈值确定为 3。

以往研究通过度中心性、节点数量等网络测度指标[33]或是其他综合指标[34]识别核心节点作为主题标签。由于知识元引用网络中节点类型相同且重复度较高，故在设定阈值后将知识元占比较高的一个或多个知识作为该群落的知识标签，以群落M_t中某个知识元K_i为例，知识标签L_{M_t}的计算过程为：

$$p_{K_i}=\frac{m_i}{\sum_{i=1}^{n} m_i} \tag{1}$$

$$p'_{K_i}=\frac{p_{K_i}-p_{\min}}{p_{\max}-p_{\min}} \tag{2}$$

$$L_{M_t}=\{K_i|p'_{K_i}>\theta\} \tag{3}$$

其中m_i为K_i在群落中的数量，$\sum_{i=1}^{n} m_i$为群落知识元总量。之后为网络内不同群落统一量纲，对各个时刻群落的p_{K_i}进行标准化处理，p_{K_i}与p'_{K_i}分别表示标准化前后K_i对应的值，$p_{\min}$与$p_{\max}$代表了标准化前群落M_t中最小与最大的值。最后得到群落M_t中大于阈值的知识标签集L_{M_t}。

2.3 领域知识群落多样性测度

知识群落与生物群落具有一定共性，本文借鉴生物多样性衡量的三个基本维度——丰富度（Variety）、均衡性（Balance）与差异度（Disparity）对领域多样性展开测度分析。丰富度用以衡量当前细分领域知识群落的种类多样性，生物学常用 Chao 指数（即群落数量）来测度丰富度V。均衡性反映了细分领域内科学研究对不同知识群落的倾向程度，受到广泛认可的知识群落往往累积更多的知识元数量，本文利用分组计算法，使用社群内数量占比计算基尼系数即均衡性B。差异度是对细分领域内不同知识群落组成的差异程度的评估，差异程度的变化彰显了当前知识集的适用性情况，生物学常采用 β 多样性分析衡量群落间差异，本文将细分领域内群落组成映射为向量，并通过计算两两间相似度得到距离矩阵，计算领域差异程度D。上述计算公式如下：

$$V=n \tag{4}$$

$$B=1-\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{1}^{i-1} p_i+\sum_{1}^{i-1} p_i}{\sum_{i=1}^{n} p_i} \tag{5}$$

chinaXiv:202304.00947v1

$$p_i = \frac{u_i}{\sum_{i=1}^{n} u_i} \quad (6)$$

$$\mathrm{D} = 1 - |\boldsymbol{S}_{ij}| \quad (7)$$

$$s_{ij} = \cos(\vec{y_i}, \vec{y_j}) \quad (8)$$

其中 n 为领域内群落数量,u_i为第 i 个群落的节点数量，$\vec{y_i}$与$\vec{y_j}$分别为第 i 和 j 个群落的知识元类型组成向量，包含知识元种类与数量，s_{ij}为群落 i 与 j 之间组成的余弦相似度，$\boldsymbol{S}_{ij}$为由s_{ij}组成的距离矩阵。类似于生物群落多样性的变化，知识群落多样性的提高初步体现为知识群落的丰富度增加、均衡性的提高以及差异度的提高，现实知识系统往往在演进中发生多个趋势的多样性变化，下文将从具体领域对知识群落演化特征进行实证分析。

3 细分领域知识群落演化实证分析

3.1 领域选择

受制于数据、工具软件与学科信息学发展程度的影响[35]，基于 SPO 的知识发现研究目前主要集中于生物医学领域，该领域具有复杂的语义关系，经由文献提取的三元组较之于关键词具有更丰富的语义表达能力。本研究使用由美国国立医学图书馆开发的 SemMedDB 知识库数据（versions 43），该数据库使用医学知识抽取工具 SemRep 对美国医学文献数据库 PubMed 抽取标题和摘要进而形成知识三元组。SemRep 按照 21 种谓词关系识别句子主语和宾语，每种谓词代表了医学研究的某一类知识三元组集合[37]，如 DIAGNOSES 代表了疾病的诊断研究、TREATS 代表了疾病的治疗研究。

SemMedDB 的核心模式描述了 SPO 三元组基本属性，涉及主语和宾语的语义类型、三元组所属文献的 pmid 号，同时关联了三元组抽取位置、文献出版时间等必要的数据项[37]（如图 2）。由于 SemMedDB 建立于 PubMed 数据库的基础之上，其 pmid 号与 PubMed 数据库中包含的文献间引用关系相对应，从而能够获取三元组间引用关系。

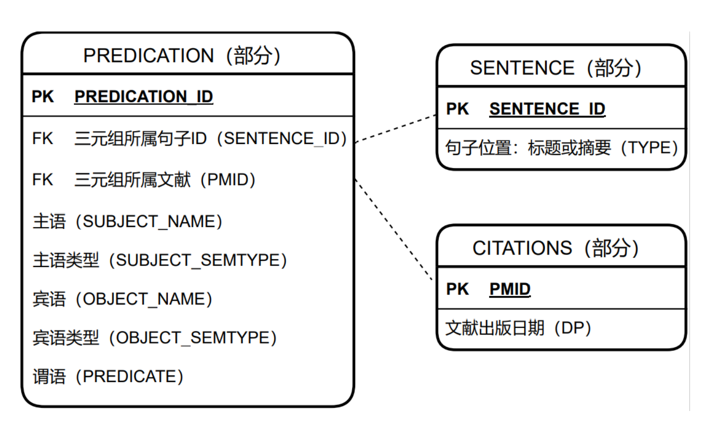


图 2 数据项描述示例

对 SemMedDB 基本信息进行统计，发现 TREATS 领域尤其是与疾病相关的

chinaXiv:202304.00947v1

治疗领域是生物医学关注的热点，占 SPO 总频次的 28.7%（见表 1）。与之相关的语义模式涉及疾病的治疗药物、治疗措施以及治疗设备等知识，即 phsu-TREATS-dsyn、topp-TREATS-dsyn、horm-TREATS-dsyn 三类语义模式。因此，本文选定 TREATS 领域中的疾病研究作为知识群落演化分析的范围。

表 1 谓词领域与 TREATS 领域语义模式 SPO 频次 top10

谓词领域	SPO 频次	语义模式	SPO 频次
TREATS	10552632	topp-TREATS-podg	1525719
AFFECTS	5172022	topp-TREATS-dsyn	1152226
COEXISTS_WITH	4342036	topp-TREATS-neop	812378
INTERACTS_WITH	4011706	phsu-TREATS-dsyn	776897
CAUSES	3140324	phsu-TREATS-podg	622437
ASSOCIATED_WITH	2695277	hlca-TREATS-humn	387540
STIMULATES	2185713	phsu-TREATS-mamm	385541
ADMINISTERED_TO	1783801	topp-TREATS-fndg	225790
INHIBITS	1590886	hlca-TREATS-dsyn	221523
AUGMENTS	1299988	topp-TREATS-popg	209580

备注：topp：治疗或预防措施；podg：病患群体；dsyn：疾病或症状；neop：肿瘤形成过程；phsu：药理物质；hlca：保健活动；humn：人群；mamn：哺乳动物；fndg：发现

3.2 细分领域数据收集与预处理

本研究首先以 TREATS 为谓语（PREDICATE）检索词在 SemMedDB 中进行初步筛选。以往研究局限于对某一类疾病的知识发现问题开展实证分析，为得到知识群落更为客观的演化特征，本文将数据收集范围扩散至研究关注度较高的多种疾病。具体操作为筛选 SPO 三元组中涉及 dsyn 的语义模式，得到 125 项数量最多的疾病名称，共包含 1048577 条三元组。

对筛选后的数据进行预处理，具体流程如下：

①删除由标题句抽取而来的三元组，其 TYPE 属性为 ti。

②清除 SUBJECT_SEMTYPE 属性为人群（humn）等不表述实际信息的三元组，包含 Physicians、Author 等。

③按照 SEBJECT_NAME 属性对不同疾病中高频出现的较为宽泛的描述治疗药物、治疗措施以及治疗设备的词语进行删除，包括 Pharmaceutical Preparations、Therapeutic procedure 等。

④将来源于相同文献的多个 SPO 三元组的 pmid 号进行去重。

最后按照节点所属文献出版时间与疾病名称分别进行时序分类与疾病分类，得到截至不同时间和疾病的节点集合，利用三元组的 pmid 号之间的引用关系构建有向边并得到边集合。对于任一疾病，选择 1960 至 2019 年期间与该疾病相关的三元组，每三年构建累积的动态知识元引用网络。

3.3 群落竞争视角下的细分领域演化状态分析

在细分领域内知识元由相同类型的知识模因组成，TREATS 领域中构建的知识元引用网络覆盖了适用于某疾病的全部治疗知识。在每一网络内的不同知识群落之间存在恒定的竞争关系，而某一群落知识标签集中的多个知识处于共生关系。对某疾病网络进行群落识别与知识标签表示后，研究利用不同时间知识标签的变

chinaXiv:202304.00947v1

化表征知识群落的演变，并利用群落组成的余弦相似度作为 t 与 t+1 时刻群落之间相似度。定义 t 与 t+1 时刻群落之间相似度为$\mathrm{Sim}(\mathrm{M}_t, \mathrm{M}_{t+1})$，其中$\overrightarrow{M_t}$与$\overrightarrow{M_{t+1}}$分别表示 t 与 t+1 时刻群落的知识元组成，其计算公式为：

$$\mathrm{Sim}(\mathrm{M}_t, \mathrm{M}_{t+1}) = \cos\left(\overrightarrow{M_t}, \ \overrightarrow{M_{t+1}}\right) \tag{9}$$

在得到不同时刻知识群落间相似度后，设定阈值筛选出知识群落之间的演化关系并构建演化路径，低于阈值的知识群落认为不存在显著的演化关系，最后对路径进行可视化。有学者将网络中社区的演化定义为新生、成长、合并、衰减、分裂以及衰亡六种模式[38]，研究结合信息的生命周期理论[39]对知识群落竞争视角下的演化状态进行定义（如表 2）。

表 2 群落竞争视角下的演化状态定义

演化状态	定义
知识产生	t+1 时刻出现的群落知识标签与 t 时刻及以前均不同
知识遗传	t 时刻与 t+1 时刻具有相同的群落知识标签
知识合并	t 时刻不同的知识群落在 t+1 时刻融入新的知识群落
知识分裂	t 时刻知识群落在 t+1 时刻分化为两个不同的知识群落
知识替代	t+1 时刻群落知识标签与 t 时刻不同
知识过时	t+1 时刻及以后不再出现 t 时刻某群落知识标签

4 结果分析

4.1 疾病治疗领域知识群落识别结果

构建 TREATS 领域（以下简称 TREATS）中 125 种疾病的不同时刻的知识元引用网络，得到 20 个时间戳下共计 2032 个网络，之后识别网络中知识群落并统计相关变量。由于自 1992 年起绝大部分疾病均已形成群落，研究将 1992 年、2004 年与 2016 年作为时间切片，呈现不同时期疾病的群落数量、群落的平均大小、网络平均长度以及引用关系数量的密度分布情况（如图 3）。

疾病的知识群落数量与规模随时间逐渐增大，并且不同疾病的群落差异也逐渐增大。从疾病群落数量的分布可以看出，历年疾病群落数量大多集中于 0 至 200 之间，而随着知识的积累，群落数量向 100 至 200 区间增加，同时 2016 年群落的平均大小整体高于 1992 年 10 个节点的群落平均大小，2016 年不同疾病的平均群落数量与大小分布更加均匀。除此之外，知识元引用网络的深度与密集程度逐渐增加。每过十年，几乎所有疾病的网络引用关系数量较之于前一阶段都会增加将近一个数量级。并且随着引用关系的累积和节点数量的增加，网络的平均路径长度分布也更加均匀且数量明显上升。

初步的群落识别结果表明，疾病的知识元引用网络随时间更加密集并且网络内的群落数量以及内部规模稳定增加，这为领域的多样化分析提供了一定依据；而不同疾病逐渐扩大的群落发展程度差异也表明，群落多样性演化可能受到其他因素的影响。

chinaXiv:202304.00947v1

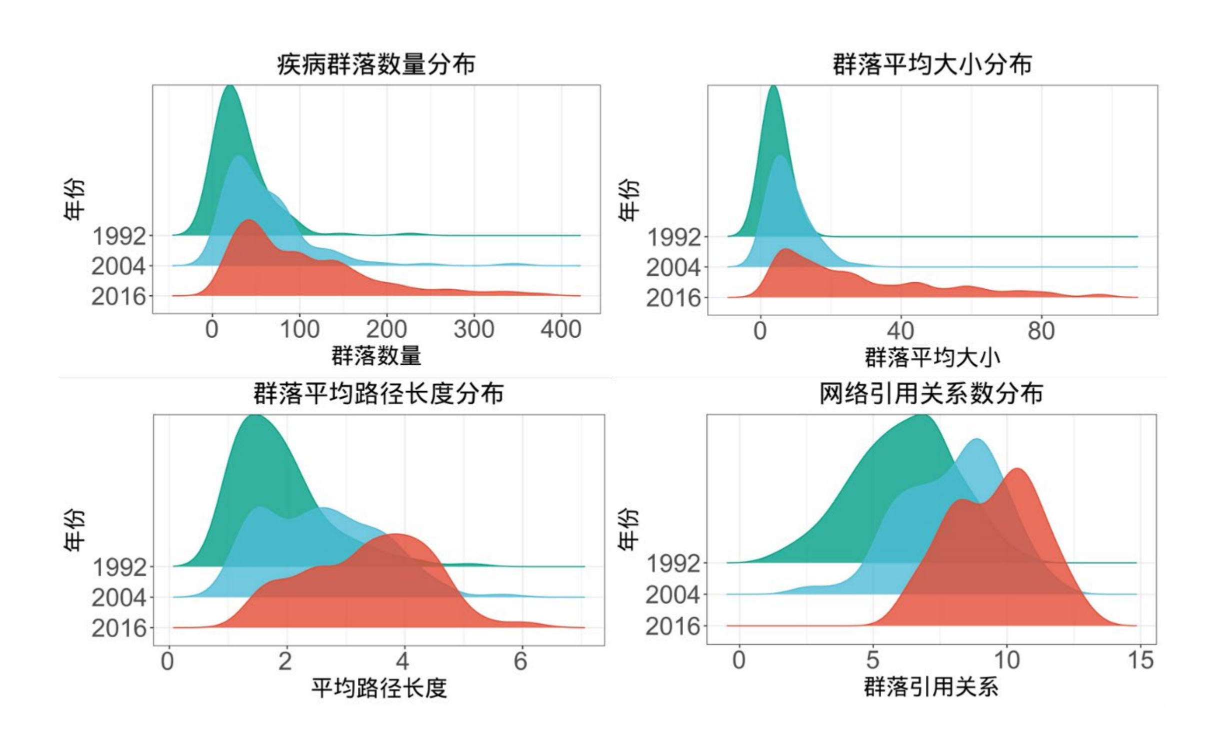


图 3 知识群落识别结果分布

4.2 示例细分领域的群落演化分析

阿尔兹海默症的疾病治疗研究最早出现于 1980 年，而后积累了大量知识并且至今仍是尚未完全攻克的研究热点之一，具有典型的分析价值。本研究以阿尔兹海默症为示例刻画并分析知识标签及群落知识演进状态，最后阐释疾病的多样性演化特征。该分析流程同样适用于其他疾病细分领域。

4.2.1 群落知识标签演化分析

本研究在表示各个时期的群落知识标签后，筛选出保持十年以上存在于较大知识群落且规模持续扩大的知识共计 10 种，统计其在网络中全部所属知识群落中的数量并绘制河流图（如图 4）。

上述知识可视为阿尔兹海默症的主流治疗方案，其大致可划分为四个阶段。疾病研究伊始毒扁豆碱（Physostigmine）在有限的知识群落内获得了绝对的关注度。从 1989 年开始他克林（Tacrine）快速扩张逐渐替代毒扁豆碱；同时非甾体抗炎药（Anti-inflammatory agents, Non-Steroidal）与司来吉兰（Selegiline）也形成规模较大的知识群落。直到 1998 年随着研究的井喷式发展，差异化的知识在不同群落中形成，早期知识群落规模逐渐停滞。这一时期，乙酰胆碱酯酶抑制剂（Acetylcholinesterase Inhibitors）、多奈哌齐（Donepezil）、卡巴拉汀（rivastigmine）以及加兰他敏（Galantamine）代替了前序知识，免疫治疗（Immunotherapy）与美金刚（Memantine）也逐渐获得研究关注。2010 年至 2019 年期间，免疫治疗与谷氨酸酯抑制剂获得更多认同，而卡巴拉汀则在知识标签演化过程中消失。同时，这一时期姜黄素（Curcumin）、深部脑刺激（Deep Brain Stimulation）、针灸（Acupuncture procedure）等新型药物或手段虽限于年份不足未被统计在内，但同样形成了较大的知识群落。

chinaXiv:202304.00947v1

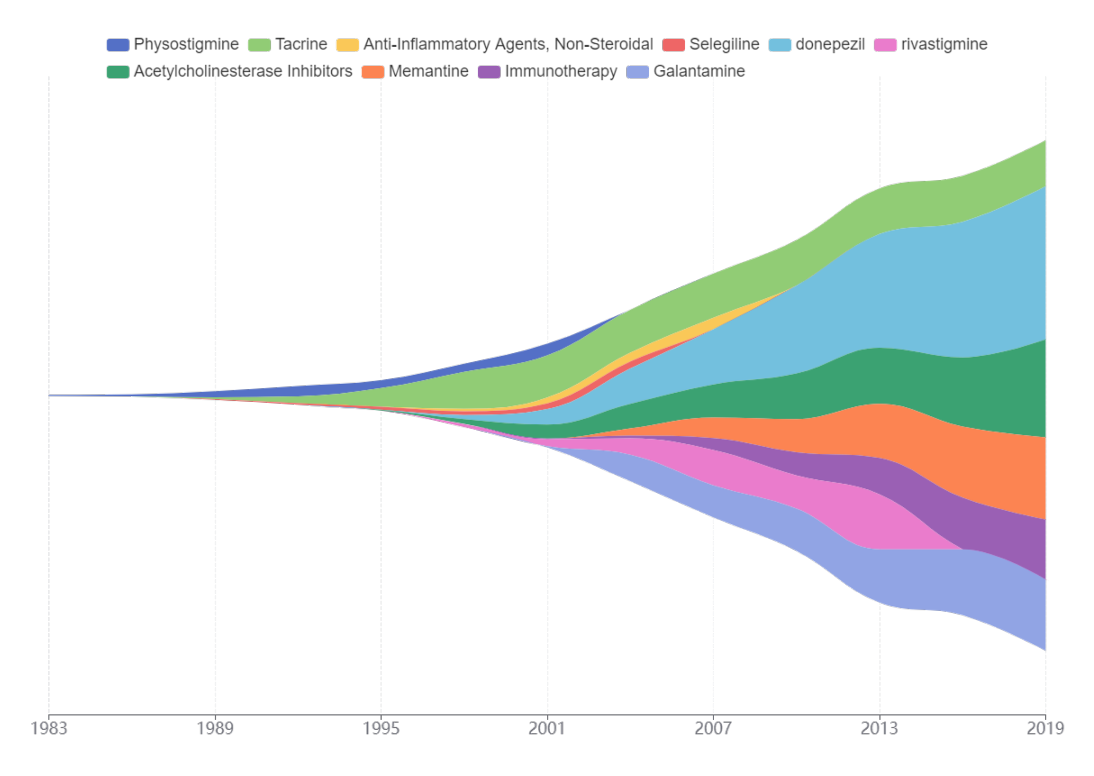


图 4 阿尔兹海默症知识标签河流图

4.2.2 群落知识竞争状态分析

以往演化研究将社区主题间的相似度阈值定为 0.7，考虑到 TREATS 内知识元种类重复度较高，充分试验后将阿尔兹海默症的知识群落相似度阈值设定为 0.8，并局部呈现包含上述主流知识的标签的路径。

本研究发现，计算知识标签后阿尔兹海默症的 10 种主流知识在各自时期均存在竞争关系。随着研究的发展，愈加多样的知识使得竞争不断加深，而知识共生关系仅存在于部分知识的合并与分裂过程。依据本文对知识演进状态的定义，以 2004 至 2016 年包含多奈哌齐与美金刚的知识标签为例，节点颜色表示不同的年份（如图 5）。自 2007 年多奈哌齐初次进行知识分裂后，一个群落发生知识遗传并于 2013 年继续发生分裂，另一个群落则经过逐步的知识替代与美金刚的另一群落发生知识合并进而维持后者的遗传状态。

归纳其他知识路径后本研究发现，新知识的产生基本从依附于当前主流知识群落到独立产生新的知识群落，而非逐步替代原有的知识群落。在主流知识的生长过程中，知识遗传是以卡巴拉汀等 6 种知识为代表的单一知识标签最为典型的特征。知识的过时意味着知识标签的消失或者群落规模的停滞，前者源于知识替代（毒扁豆碱）或是群落的重组（卡巴拉汀），后者表现为相邻时刻群落相似度接近于 1，这一点在 2004 年前后他克林的群落相似度表现得到了初步证明，而美金刚在 2013 至 2019 年均维持在 0.99 以上的相似度意味着其将可能面临知识过时。

局限于 SemRep 工具的抽取结果，部分知识标签仍然存在概念较为相似的问题，例如胆碱酯酶抑制剂（Cholinesterase Inhibitors）与卡巴拉汀等药物均存在药理上的重合。限于研究篇幅，研究对大范围的知识演化机制并未进行深入探究，但通过知识标签追踪主流知识的演化过程，为刻画知识生命周期提供了可行的路径。

chinaXiv:202304.00947v1

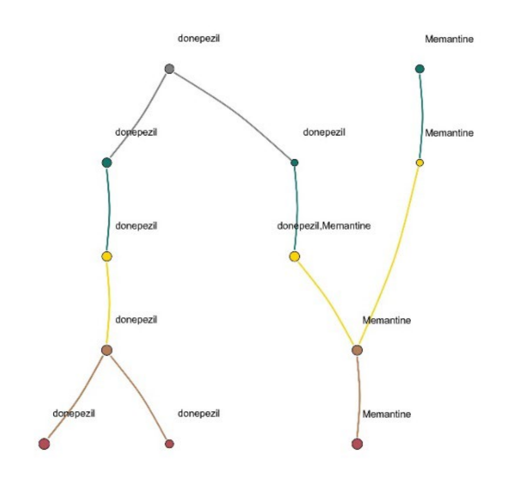


图 5 知识演进局部路径

chinaXiv:202304.00947v1

4.2.3 群落多样性特征分析

群落的多样性演化指标表明，阿尔兹海默症的知识群落早期增速较快，且丰富度规模持续增加，在 2019 年达到 63 个群落。从群落均衡性上看，疾病研究早期存在较高的基尼系数跨度，均衡性大幅降低，结合知识标签演化表明这一时期他克林在有限的知识中得到大量积累，而在 2000 年后大致维持在 0.8 左右的均衡值。疾病研究则保持着始终较高的差异度，早期由于知识群落数量较少且知识元种类较为单一，群落间差异度接近于 1，之后群落差异度波动下降并稳定 0.9 左右（如图 6）。

生物多样性理论认为，较高的丰富度、均衡性与差异度一般意味着较高的领域多样性，也即是阿尔兹海默症研究在均衡性指标上结果呈现不佳。但对于疾病治疗领域，缺乏较大的独占性群落意味着细分领域尚未形成受到主导性认可的知识，表明对该疾病的研究尚未找到一个共识性的解决方案，因而知识群落多样性特征的分析还需结合细分领域的具体语义。结合上文知识标签演化与竞争状态分析，在疾病研究逐渐成熟的过程中，阿尔兹海默症的知识标签种类逐渐增加，并且伴随着知识群落丰富度与差异度的持续增长。

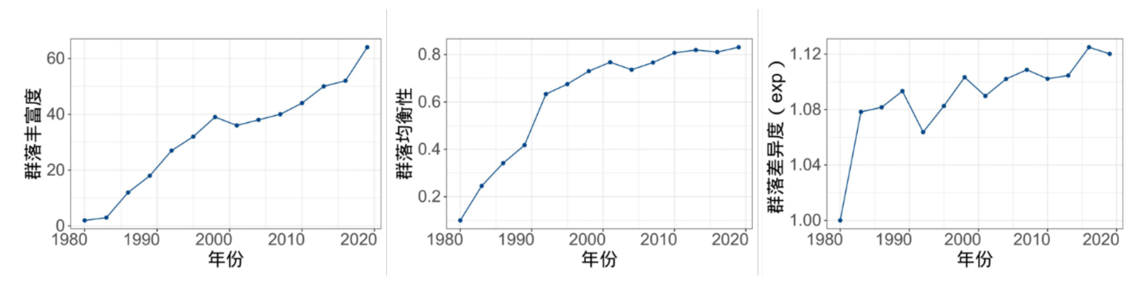


图 6 阿尔兹海默症多样性特征

4.3 整体研究领域的多样性演化特征

除阿尔兹海默症外，本研究测度了所选 125 种疾病的多样性演化特征，并分析其相似或相异的多样性表现。

4.3.1 多样性演化一般性特征

本研究首先统计了 125 种疾病的三项多样性指标的平均值与方差，对拟合后的结果分析疾病多样性演化特征的共性（如图 7）。

TREATS 中疾病丰富度逐渐提高，且平均群落规模的增速接近于二次函数型增长（红色曲线），但方差的指数型增长表明疾病间群落规模的差异显著增大，表明即使是包含 SPO 三元组数量最多的 125 种疾病，研究知识仍存在较大的富集差异。此外，知识群落的平均基尼系数在统计年限内大致呈现 S 型增长，说明知识演化系统的领域均衡性不断下降且中期改变幅度较大。2019 年 TREATS 内较多疾病出现了少数占据绝对支配地位的知识群落，意味着组成该群落的药物知识在疾病治疗效力上得到了高度认同。均衡性方差在 1992 年后持续下降证明了疾病间均衡性差异逐渐缩小，部分疾病基尼系数逐渐稳定。

最后疾病的群落差异度随时间小幅下降但总体较高，知识群落保持着较大的差异程度。差异性方差的波动下降，表明疾病内知识群落的组成差异变化更加相似。进一步统计 125 种疾病的差异度数据，发现其分化为大幅升高与小幅较低两种相异趋势，且 2019 年大部分疾病的相似度稳定在 0.09 附近，结合示例疾病推断低相似度的表现可能来源于知识标签的多样化以及群落竞争状态的深化。

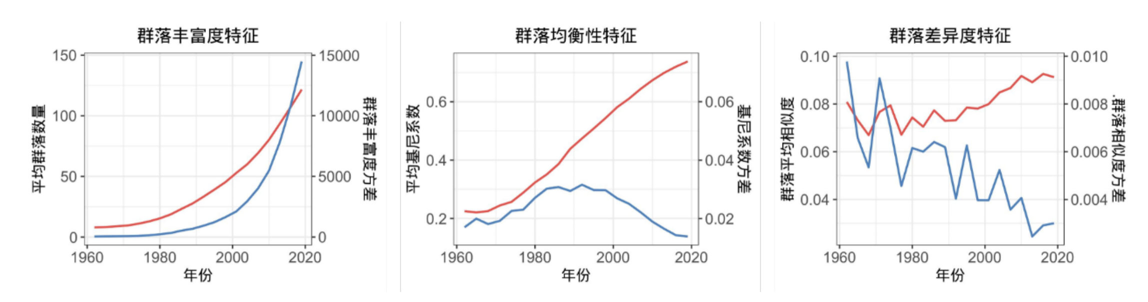


图 7 整体领域疾病多样性演变相似特征

4.3.2 多样性演化区位特征

丰富度、均衡性以及差异度指标的综合呈现能够体现不同疾病的区位分布情况。研究分别以横轴、纵轴代表均衡性与差异度指标，点的大小表示疾病群落丰富度规模，点的颜色代表不同的年份，得到 125 种疾病的散点图（如图 8）。从 1962 年至 2019 年颜色的分布可以发现疾病的演化大致遵从以下三种模式：一类是由低差异度高均衡性的小规模疾病研究向高差异度低均衡性的大规模研究迈进；后两类是由高差异度高均衡性的小规模疾病分别向中等均衡性与低均衡性的大规模研究演进，且差异度小幅下降。

chinaXiv:202304.00947v1

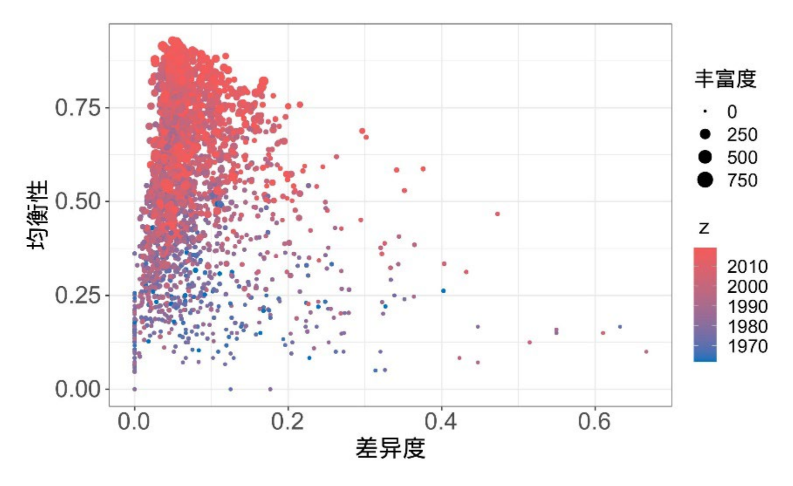


图 8 整体领域疾病多样性演变相异特征

不同疾病的发展模式体现出内部差异化的知识演进过程。疾病知识群落规模的稳定增长表明细分领域知识丰富度持续增加。但差异度大幅提高的演进模式表明疾病的早期研究阶段治疗药物较为单一，知识群落组成随着研究深入日趋复杂，疾病治疗逐渐多元化，可将其视作"常规型"演进模式。反之疾病研究早期的高差异度可能源于出现了几类具有争议性的治疗方案且只在各自的群落里认可，而随着疾病研究的发展，某一类药物或是新药物开始占据疾病研究的主体地位，本研究将向着高差异度低均衡性演进的疾病视为"早期争议型"模式。最后部分疾病在 2019 年表现的中等均衡性表明其尚未出现占据绝对支配地位的知识群落，为初步探究这一现象的成因，本文统计了部分中等均衡与低均衡疾病的名称及其在 2019 年的基尼系数（见表 3），可以发现中等均衡性的疾病概念较为抽象，多为某一身体系统或器官的通用性描述；低均衡性特征则更多指向专指性疾病，本研究将高差异度中等均衡性疾病视作"泛指型"知识演进模式。

表 3 低均衡与中等均衡疾病描述差异

低均衡性疾病	基尼系数	中等均衡性疾病	基尼系数
II型糖尿病	0.928	血管疾病	0.436
高血压疾病	0.924	传染性疾病	0.437
哮喘	0.908	心脏疾病	0.477
急性心肌梗死	0.903	疹类疾病	0.519
心力衰竭	0.893	病毒性疾病	0.523
糖尿病	0.893	神经系统疾病	0.528
败血症	0.889	肾衰竭	0.536
克罗恩病	0.888	非酒精性脂肪肝	0.541
结核病	0.887	多囊卵巢综合征	0.541
脑血管病变	0.882	乳糜泻	0.558

4.3.3 多样性演化时间特征

考虑到疾病的研究起始时间这一因素，本研究探究其与疾病知识群落多样性发展是否有关。研究所取疾病自起始研究时间开始未曾中断，通过获取疾病样本的时间戳数量可粗略得到该疾病的研究时长。对不同研究时长的疾病进行聚类并

chinaXiv:202304.00947v1

计算多样性指标的平均值，得到疾病研究时间差异气泡图（如图 9）。纵轴自下而上研究时间依次增加，横轴表示基尼系数，颜色深浅表示相似度大小，点的大小表示群落数量。

虽然起始研究时间较早的疾病数量远高于短期研究疾病，但结果仍可以表明，早期开始研究的疾病具有更高的基尼系数与更低的相似度。结合前文多样性演变的一般规律，说明起始时间较早的研究形成了更高的差异度与更低的均衡性。但不同起始时间的疾病并未形成较大的丰富度差异，表明知识群落规模的增长并不完全受到起始时间因素的制约，一些新出现的疾病可能会在短时间内受到更多的关注与研究，从而快速积累群落规模。

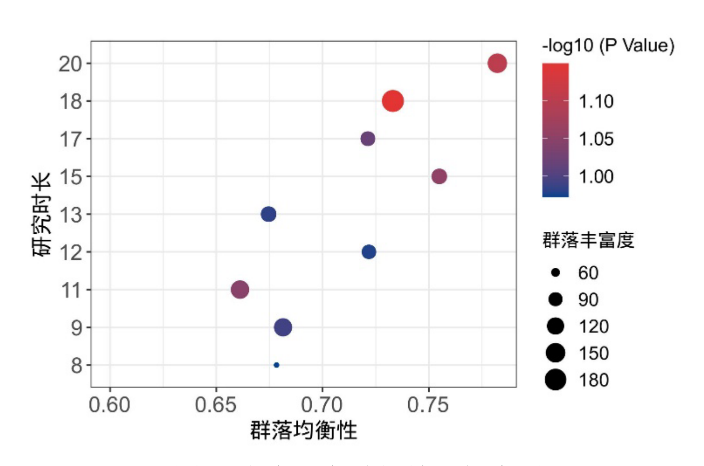


图 9 疾病研究时长差异气泡图

5 总结与展望

本文以知识元三元组为研究对象，提出知识元引用网络并设计细分领域知识群落演化的分析框架与研究流程，在借鉴丰富度、均衡性与差异性三种多样性测度指标基础上对生物医学领域展开了较为全面的多样性演化分析，本文主要得出以下结论。

（1）方法丰富了知识群落的相关研究，通过示例疾病的群落演化验证了分析知识演化路径以及演进状态的可行性。本研究梳理了阿尔兹海默症四个阶段的 10 种主流治疗方案，并分析其生命周期中的遗传、替代等知识演进状态，最后结合疾病在均衡性上的表现提出知识多样性的评估应结合具体的语义内涵。

（2）本文发现与治疗相关的生物医学领域遵从三个方面的多样性演化规律。一是领域内疾病丰富度呈现二次函数型增长且均衡性呈现 S 型降低，二是根据多样性指标的综合呈现可大致划分出常规型、早期争议型以及泛指型三类疾病演化模式，三是开展研究较早的疾病倾向于具有更高的差异度与更低的均衡性。

本文仍存在一定不足。例如并未考虑知识元在文章中的语义功能，未来可以融合方法、结论等功能为知识元附加更多可供挖掘的信息。除此之外，实证分析未深入到疾病其他领域知识的关联，对疾病的外源性信息结合不够紧密，无法做出更深入的医学解释。如何更具目的性的抽取 SPO 三元组为并系统性地构建路径将会是未来研究的重点。

参考文献:

[1] 孙震, 冷伏海. 一种基于知识元共现的 ESI 研究前沿知识演进分析方法[J].

chinaXiv:202304.00947v1

chinaXiv:202304.00947v1

情报学报, 2018, 37(11): 1095-1113.
[2] ZHANG R, HRISTOVSKI D, SCHUTTE D, et al. Drug repurposing for COVID-19 via knowledge graph completion[J]. Journal of Biomedical Informatics, 2021, 115: 103696-103696.
[3] SONG M, HAN N, KIM Y, et al. Discovering implicit entity relation with the gene-citation-gene network[J]. PloS one, 2013, 8(12): e84639.
[4] Du J, LI X. A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development[J]. JMIR medical informatics, 2020, 8(4): e18323.
[5] 杜建. 医学知识不确定性测度的进展与展望[J]. 数据分析与知识发现, 2020, 4(10): 14-27.
[6] 温有奎. 基于"知识元"的知识组织与检索[J]. 计算机工程与应用, 2005, (01): 55-57+91.
[7] 索传军, 盖双双. 知识元的内涵、结构与描述模型研究[J]. 中国图书馆学报, 2018, 44(04): 54-72.
[8] 王雪, 杨雪梅, 李沛鑫, 等. 基于语义模型的药物矛盾知识发现[J]. 情报杂志, 2020, 39(07): 159-165.
[9] 蔡妙芝, 李晓瑛, 赵嘉玮, 等. 基于SPO语义三元组的疾病知识发现[J]. 数据分析与知识发现, 2022, 6(01): 134-144.
[10] KUHN T, PERC M, HELBING D. Inheritance patterns in citation networks reveal scientific memes[J]. Physical Review X, 2014, 4(4): 041036.
[11] MAO J, LIANG Z, CAO Y, et al. Quantifying cross-disciplinary knowledge flow from the perspective of content: Introducing an approach based on knowledge memes[J]. Journal of Informetrics, 2020, 14(4): 1751-1577.
[12] 操玉杰, 梁镇涛, 毛进. 知识模因视角下跨学科研究领域的学科结构分析[J]. 图书馆论坛, 2019, 39(07): 84-90.
[13] K·波普尔. 客观知识——一个进化论的研究[M]. 舒炜光等, 译. 上海: 上海译文出版社, 1987: 114.
[14] 王曰芬, 丁玉飞. 基于知识进化视角的科学文献传播网络演变模型构建及仿真[J]. 情报学报, 2019, 38(09): 966-973.
[15] 滕广青, 杨明秋, 田依林, 等. Folksonomy 模式中的知识群落及其核心知识分析[J]. 图书情报工作, 2015, 59(22): 124-129.
[16] PHAM M C, KLAMMA R, JARKE M. Development of computer science disciplines: a social network analysis approach[J]. Social Network Analysis and Mining, 2011, 1(4): 321–340.
[17] 滕广青, 贺德方, 彭洁, 等. 基于网络演化的领域知识群落生长机制研究[J]. 情报理论与实践, 2016, 39(10): 16-20+15.
[18] XU J, DING Y, BU Y, et al. Interdisciplinary scholarly communication: An exploratory study for the field of joint attention[J]. Scientometrics, 2019, 119(3): 1597–1619.
[19] 马铭, 王超, 周勇, 等. 基于语义信息的核心技术主题识别与演化趋势分析方法研究[J]. 情报理论与实践, 2021, 44(09): 106-113.
[20] 靳嘉林, 王曰芬, 巴志超, 等. 基金项目研究的主题挖掘与动态演化分析——以美国 NSF 数据中 AI 领域为例[J]. 情报学报, 2022, 41(09): 967-979.

[21] HU Z, ZENG R, PENG L, et al. Discovering Emerging Research Topics Based on SPO Predications[C]//L, Uden et al. Communications in Computer and Information Science. Zamora, Spain: Springer, 2019: 110-121.
[22] 陈翔，黄璐，倪兴兴，等．基于动态语义网络分析的主题演化路径识别研究[J]．情报学报, 2021, 40(05): 500-512.
[23] 黄萃，黄施旗，付慧真．学科交叉视角下人工智能治理领域知识流动与研究主题的国际比较研究[J]．信息资源管理学报,2022, 12(06): 98-110.
[24] 胡吉明，杨泽贤，朱国伟，等．中国政府在做什么——基于词汇关联的《政府工作报告》内容结构分析[J]．信息资源管理学报, 2023, 13(01): 115-128.
[25] 李维思，谭力铭，章国亮，等．基于多源信息融合的产业链关键核心技术主题识别研究——以人工智能领域为例[J]．信息资源管理学报, 2022, 12(01): 116-126.
[26] WANG X, HE J, HUANG H, et al. MatrixSim: A new method for detecting the evolution paths of research topics[J]. Journal of Informetrics, 2022, 16(4): 1751-1577.
[27] KIM E, JEONG Y K, KIM Y H, et al, Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction[J]. Journal of Informetrics, 2022, 16(1): 1751-1577.
[28] ZHANG X, XIE Q, SONG C, et al. Mining the evolutionary process of knowledge through multiple relationships between keywords[J]. Scientometrics, 2022, 127(4): 2023–2053.
[29] 梁镇涛，毛进，操玉杰，等．基于知识模因级联网络的领域知识扩散模式分析[J]．情报理论与实践, 2020, 43(04): 40-46+39.
[30] 白如江，孙一钢，张庆芝．基于知识基因表达的科技创新路径构建研究[J]．情报理论与实践, 2020, 43(04): 137-144+176.
[31] 孙震，冷伏海．一种基于知识元迁移的 ESI 研究前沿知识演进分析方法[J]．情报学报, 2021, 40(10): 1027-1042.
[32] TRAAG V A, WALTMAN L, VAN ECK N J, From Louvain to Leiden: guaranteeing well-connected communities[J]. Scientific Reports, 2019, 9(1): 2045-2322
[33] 赫南，李德毅，淦文燕，等．复杂网络中重要性节点发掘综述[J]．计算机科学, 2007(12): 1-5+17.
[34] 安世虎，聂培尧，贺国光．节点赋权网络中节点重要性的综合测度法[J]．管理科学学报, 2006(06): 37-42+52.
[35] 代冰，胡正银．基于文献的知识发现新近研究综述[J]．数据分析与知识发现, 2021, 5(04): 1-12.
[36] KILICOGLU H, ROSEMBLAT G, FISZMAN M, et al. Broad-coverage biomedical relation extraction with SemRep[J]. BMC Bioinformatics, 2020, 21(1):188.
[37] KILICOGLU H, SHIN D, FISZMAN M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications[J]. Bioinformatics, 2012, 28(23): 3158–3160.
[38] PALLA G, BARABÁSI A L, VICSEK, T. Quantifying social group evolution[J]. Nature, 2007, 446(7136): 664–667.
[39] 索传军．试论信息生命周期的概念及研究内容[J]．图书情报工作, 2010, 54(13): 5-9.

chinaXiv:202304.00947v1

作者贡献说明：毛进：提出论文研究思路、修订论文；侯博文：文献调研、数据整理与分析、撰写与修订论文；王依蒙：数据收集与整理。

Research on the Evolution Characteristics of Subdivision Fields based on Knowledge Unit Citation Networks

Mao Jin[1,2] Hou Bowen[1,2] Wang Yimeng[2]
[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2] School of Information Management, Wuhan University, Wuhan 430072

Abstract: **[Purpose/Significance]**Comprehending the development process of scientific knowledge contributes to the scientific research.. For knowledge evaluation and service, it is crucial to trace the structure and progress of knowledge in subdivided fields from a micro perspective. **[Method/Process]**This article took the knowledge unit in medical informatics as an example. This paper used the semantic type of triples to define the treatment-related subdivision fields, constructed the knowledge unit citation networks of 125 diseases at different intervals, and identified the knowledge communities with the Leiden algorithm. From the dimensions of knowledge evolution and knowledge competition state, we aim to reveal the evolutionary characteristics of disease individuals. The indexes of Richness, Balance, and Difference are calculated to reveal the diversity characteristics of disease individuals as well as the overall disease population. **[Result/conclusion]**The research demonstrates that the knowledge communities can reflect the knowledge structure and evolution state of disease individuals. The overall diversity characteristics of diseases include: the commonality of indicators indicates that the number of all disease knowledge communities is increasing, and the differences in scale and composition between communities are expanding. Different diseases show conventional, early-controversial, and generalized evolutionary patterns, with the earlier diseases being less balanced and more different.
Keywords: SPO triples knowledge unit knowledge community evolution characteristics knowledge diversity

chinaXiv:202304.00947v1